RETHINKING EMPIRICAL EVIDENCE OF DISCRIMINANT VALIDITY: THE ROLE OF CONFIRMATORY FACTOR ANALYSIS IN CONSTRUCT PROLIFERATION

by

Betsy H. Albritton

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Arts in
Industrial Organizational Psychology

Charlotte

2021

Approved by:

_____
Dr. Scott Tonidandel

_____
Dr. Dave Woehr

_____
Dr. Eric Heggestad

ABSTRACT

BETSY H. ALBRITTON. Rethinking Empirical Evidence Of Discriminant Validity: The Role Of Confirmatory Factor Analysis In Construct Proliferation
(Under the direction of DR. SCOTT TONIDANDEL)

Our field acknowledged the importance of discriminant validity since Campbell and Fiske's (1959) multitrait-multimethod matrices (MTMM), and yet, the presence of construct proliferation serves as evidence of the lack of clarity or insufficiency of current practices for justifying discriminant validity inferences. This paper seeks to improve our measurement sciences and combat construct proliferation by providing better empirical guidelines for supporting discriminant validity of constructs. We present how researchers are typically supporting the discriminant validity of constructs through a systematic literature review of 849 articles from five top journals in organizational science. Additionally, using results from 325,000 iterations of a Monte Carlo simulation, we demonstrate the questionable efficacy of a specific and dominant methodological approach, confirmatory factor analysis (CFA), in identifying unitary constructs. We compare these results to existing recommended practices and seek expert opinions from the *Organizational Research Methods* (*ORM)* editorial board on the specific practices that researchers should be engaging in to test discriminant validity inferences. Finally, using this data, we provide a set of recommendations for researchers when evaluating the distinctiveness of their measures to mitigate the issue of construct proliferation moving forward.

ACKNOWLEDGEMENTS

I would like to thank my committee chair and advisor, Dr. Scott Tonidandel, for supporting and guiding me through this project. His mentorship and expertise contributed greatly to the development of this project and myself as a scholar. I am grateful for each opportunity to work with him.

I would also like to thank my committee, Drs. Dave Woehr and Eric Heggestad, for their consistent input and support throughout the duration of the project. Their advice was instrumental in the creation, development, and completion of this paper.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1: INTRODUCTION

The value of our science is dependent on good measurement, since scientific progress becomes stifled when measures become inexact. When redundant concepts are measured but treated as theoretically distinct, the corresponding theory and research incorporating those concepts is diluted and imprecise. At the heart of this problem is construct proliferation. Construct proliferation, the systemic formation and publication of multiple constructs when only one conceptual and empirical construct exists, is a pervasive problem in the organizational sciences (Le et al., 2010; Shaffer et al., 2016). As an example, consider emotional intelligence. Some studies support the distinctiveness of emotional intelligence and traits (Ashkanasy & Daus, 2005), while others argue that emotional intelligence is not separate from trait and general intelligence constructs (Antonakis, 2004; Antonakis & Dietz, 2010). The lack of clarity surrounding this issue presents barriers to our scientific advancement. When questions such as these arise, the tools we use to support construct distinctiveness become critically important to the resolution of debates and production of precise scientific progress.

The purpose of this paper is to improve our measurement sciences and combat construct proliferation by providing better empirical guidelines for establishing discriminant validity of constructs. We first catalog how people are typically supporting discriminant validity of constructs by looking at the procedures and evidence they provide to justify their decisions. Next, using a Monte Carlo simulation, we evaluate the efficacy of these methodological approaches for identifying unitary constructs and compare these results to recommended practices. We also evaluate how current practices and our simulation results align with expert opinions about how and when a measure

should be considered empirically distinct and discuss the implications of certain decisions in the empirical establishment of unitary constructs. Ultimately, we provide a set of recommendations for researchers when evaluating the distinctiveness of their measures to mitigate the issue of construct proliferation moving forward.

Validity

Constructs are underlying representations of behavioral groupings that tend to covary (Binning & Barrett, 1989). Researchers strive to define these latent behavioral groupings to describe phenomenon and expand theory. A critical process in the definition and measurement of constructs is the establishment of validity. Validity assists researchers in grouping and testing behavioral clusters to create constructs that measure what we believe them to represent. Stated differently, validity is the "degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA et al., 2014, p. 11). This broad definition of validity was originally broken down into three types of validity identified in the 1950s: criterion-related, content, and construct validity (Campbell & Fiske, 1959). Over the subsequent decades, researchers further outlined validity types beyond the core three (i.e., face validity, convergent validity, discriminant validity) until critics argued for the abolishment of the term validity "types." They asserted that validity is one core concept so there cannot be different types, but different inferences or investigations of validity (Binning & Barrett, 1989; Sackett et al., 2012).

Different tests and evidence support different validity inferences for the measures of constructs. Tests of validity require evidence from each stage of the measure development and testing. In total, the validation process requires five sources of validity

evidence: the relationship between the measure and other variables, the content of the measure, the internal structure of the measure, the response processes, and the potentially negative consequences of the measure (American Psychological Association, 2018), but certain inferences do not require all sources of validity evidence. For example, inferences of construct validity are defined as convergence with related constructs and divergence from unrelated constructs (Sackett et al., 2012); therefore, the validity evidence required to support this inference is the relationship between the measure and other variables. The tools and evidence organizational researchers use to test and support inferences of validity are vital components of our science. The present study explores the validation process and evidence of a specific facet of the construct validity inference: discriminant validity.

Discriminant validity

Discriminant validity supports the distinctiveness of a construct due to the absence of relatedness with other constructs. Tests of discriminant validity focus specifically on the relationship between the measure and other variables, but also considers the internal structure of the measure (i.e., items). Discriminant validity and the establishment of discriminant validity through various testing methods and validation processes is critical for the elimination of redundant constructs in the organizational literature; nevertheless, our current standards for establishing discriminant validity appear to be failing us.

Construct proliferation

Our models are getting more and more complex over time due to the increasingly high volume of novel constructs introduced each year. Our field has grown to include

over 8,000 published articles in over 170 journals each year in management alone (Davis, 2015). Looking deeper, there is also greater complexity in the articles themselves. Saylors and Trafimow (2021) reviewed all of the articles published in three major journals over a ten-year period. They found that the number of variables in causal models drastically increased over this period, increasing the likelihood of false models. The average number of variables in the most complex causal models was 16 during the period of 2016 to 2018, but the largest number of variables in a causal model in 2008 was 12 (Saylors & Trafimow, 2021). One contributing factor of this increase in complexity may be construct proliferation (also referred to as construct redundancy).

Construct proliferation conflicts with the goal of creating parsimonious theory and models in organizational research, but the practice of generating novel constructs and theory is heavily incentivized by our distinctive fields (Hambrick, 2007). The main currency in academia is citations, which often favors novelty over good measurement or sound scientific practices (Davis, 2015). Academic career systems require faculty to actively publish in top-tier journals and receive high numbers of citations of those publications, and the journal system publishes and rewards novel claims, but infrequently punishes questionable findings or unethical scientific practices (Davis, 2015). These incentives engender complex theory building and the proliferation of constructs in opposition to traditional parsimonious models. Although there is general agreement on the presence of construct proliferation and continuing conversation surrounding the threats of construct redundancy to scientific progress in organizational research, we lack research on methodological explanations of this phenomenon.

Methods utilized in establishing discriminant validity

There are two approaches for establishing the distinctiveness of constructs, one conceptual and one empirical. Researchers use extant literature and theory to generate hypothetical constructs. Hypothetical constructs are conceptual definitions that group characteristics and features to create representations of our real or phenomenological world (Podsakoff et al., 2016). Inherently, this grouping of features demonstrates its distinctiveness from other hypothetical constructs, because other hypothetical constructs contain different feature groupings. Conceptual distinctiveness can strengthen the empirical tests of distinctiveness via conceptual clarity and informed measurement (Podsakoff et al., 2016) and is critical to the interpretation of theory and research, but conceptual distinctiveness is not the sole source of evidence in the support of discriminant validity inferences. Constructs can be conceptually similar, but empirically dissimilar (Shaffer et al., 2016) or vice versa (Harter & Schmidt, 2008). While we recognize the importance of clear conceptual definitions in establishing discriminant validity, our study focuses on evaluating current empirical tests of redundancy and challenges their effectiveness.

To consider the root causes of the proliferation of constructs, we reflected on past and current methods for empirically establishing discriminant validity. Campbell and Fiske's (1959) seminal piece on multitrait-multimethod matrices (MTMM) defined discriminant validity as low correlations between traits when using the same method (i.e., heterotrait-monomethod). There was no singular threshold for concluding two constructs were distinct. Rather, general comparisons were made between the correlations in the validity diagonals and the correlations in the heterotrait-monomethod triangles. The MTMM matrix approach poses some difficulties in application. First, the observed

correlational magnitudes must be evaluated by researchers without clear guidelines. Second, deriving clear and consistent patterns is more challenging as the number of measures being compared increases. The range of correlational values that contest or support discriminant validity is largely unagreed upon and more sophisticated tests have been proposed (e.g. MTMM-CFA). Nevertheless, since this publication, organizational researchers have adapted a simplified version of this overall approach to consider the magnitude of a simple product-moment correlation between two potential constructs as evidence for or against discriminant validity. A high observed correlation is interpreted as covariation between two constructs indicating weak evidence of discriminant validity. Contrarily, an exceptionally low correlation would support for differences between two constructs. However, clear guidelines about what constitutes a high or low correlation remain elusive.

Modern methods of evaluating inferences of construct validity typically involve examining the fit of a confirmatory factor analysis (CFA) measurement model. Using that measurement model as the starting point, multiple indices have been developed to support the inference of discriminant validity. One such measure is the average variance extracted (AVE). The AVE estimate for a construct is the ratio of squared standardized item loadings divided by the total variance. Fornell and Larcker (1981) advocated for this approach since they believed that structural consistency is a better determinant of goodness of fit compared to traditional characteristics of the correlation matrix advocated by others (e.g. Bagozzi, 1981). To support inferences of discriminant validity, the AVE estimate should be greater than the shared variance between the constructs (Fornell &

Larcker, 1981). As evidence of the popularity of the AVE approach, the Fornell and Larcker (1981) paper has been cited 61,173 times.

Another approach consistently used to establish discriminant validity of constructs is the comparison of models in CFA. Consider a simple example where one wishes to establish the discriminant validity of two constructs. Researchers will fit both a two-factor model and single-factor model (usually by constraining the covariance between the two factors to be 1), and because these two models are nested, one can then compare the fit indices to establish discriminant validity. Researchers frequently employ a chi-square (χ^2) test of difference to compare the χ^2 values of the nested models. Researchers also may justify discriminant validity based upon the change in other commonly reported fit indices such as the Tucker-Lewis index (TLI), comparative fit index (CFI), root mean square error (RMSEA), and standardized root mean square residual (SRMR). There is no consistent standard or rule for which fit statistics are most appropriate to report as evidence of discriminant validity, but past research provides some useful guidance for when to leverage or avoid certain fit indices. In spite of its frequent use, Monte Carlo simulation studies demonstrated the sensitivity of the χ^2 test to sample size prompting researchers to explore other measures of model fit within CFA (MacCallum et al., 1999; Marsh et al., 1998; Mundfrom et al., 2005).

Beyond sample size, the number of factors, correlations between factors, and number of items per factor also contribute to model fit inaccuracies. Cheung and Rensvold (2002) considered these conditions and more in their simulation to identify a goodness-of-fit index (GFI) that demonstrated the lowest levels of sensitivity to changes in these conditions. The simulation results showed that CFI, gamma-hat, and McDonald's

NCI were not redundant or highly correlated with other fit indices and were not affected by the complexity of the models. Cheung and Rensvold (2002) concluded that CFI, gamma-hat, and McDonald's NCI performed best compared to χ^2, TLI, RMSEA, and additional GFIs. Additionally, the authors recommended a cutoff value of 0.01 for the change in CFI to avoid high incidences Type 1 error. These results offer insight into temperamentality of fit indices and the absence of consistent, reliable recommendations within the context of discriminant validity tests.

Shaffer et al. (2016) provided suggestions for empirically testing construct redundancy in the organizational science literature. They addressed the use of MTMM matrices and CFA in the context of discriminant validity and raise important concerns regarding the confidence organizational researchers can have in various empirical approaches. The authors review each step of the validation process from the establishment of conceptual distinction through a literature review to the empirical testing of discriminant validity and the correction for random, transient, or specific factor error. In their investigation of these sources of error in discriminant validity testing, Shaffer et al. (2016) emphasized the need to address these assumed standards of factor analytic approaches in testing discriminant validity inferences. Some important concerns include the considerable assumptions made by CFA that remain untested in the literature such as the heavy reliance on statistical tests such as the χ^2 difference test for two models with constrained and unconstrained covariance. Once again, the alternative to χ^2 difference tests are not much improved since they simply rely on "rules of thumb" or commonly accepted cutoffs. This study will challenge these assumptions that researchers in our field consistently make today. The present paper seeks to explore the current practices, issues

related to these practices, and offer recommendations to absolve the presence of construct redundancy in our literature.

CHATPER 2: STUDY 1 - SYSTEMATIC LITERATURE REVIEW

Given all the different approaches for supporting discriminant validity, we wanted to see how researchers justify inferences of discriminant validity and what researchers ultimately conclude as a result of their methods and justification. We conducted a systematic literature review to collect data on the common practices of organizational researchers when using empirical methods to support claims of discriminant validity. We utilized the data collected from the systematic review to define the common practices used by organizational researchers to support their claims of discriminant validity, and examined the statistical evidence that researchers used to justify their decision.

Method

Sample

We coded articles from five top-tier journals (*Journal of Applied Psychology*, *Journal of Management*, *Personnel Psychology*, *Strategic Management Journal*, and *Academy of Management Journal*) over a period of two years (2018 to 2019). We conducted a pilot literature search in the 2018 volumes of the *Journal of Applied Psychology* to identify articles that contained an empirical test of discriminant validity. Out of a total of 83 articles, 42% contained confirmatory factor analyses that tested discriminant validity or measurement models. After we identified the articles, we extracted a list of keywords commonly mentioned in the articles (Appendix B). We used the complete list of keywords in an electronic literature search to pull articles from the other journals and across the other years. The pilot literature search also helped develop the coding procedure described in the section below.

Coding procedure

To code the literature, the primary investigator trained an additional coder. This team coded a common set of articles from the 2018 issues of *JAP* ($N = 32$) and the percent agreement for twenty coding decisions was assessed. After one round of coding, acceptable inter-rater agreement was reached ($M = 99.15\%$) and any disagreements over coding were discussed. Then, each coder was assigned a sample of journals and years to code individually. The coding team compiled data through an online survey platform, Qualtrics. Aside from general information about the article, we coded for five specific pieces of data: the types of models being tested (i.e., 2-factor to 1-factor), descriptive information about the measures (i.e., sample size, number of items and reliability for each measurement scale, the average variance extracted (AVE), the observed correlation between measures, and the latent variable correlation(s)), reported global fit indices, their ultimate decision based on their reported evidence, and the main justification for their decision. The specific coding questions from Qualtrics can be found in Appendix A. Ultimately, we were evaluating which statistics and at what values authors choose to report as evidence for the discriminant validity inference.

Results

Results from the systematic literature review revealed multiple themes regarding the current practices for selecting and reporting empirical evidence of discriminant validity inferences, including the frequency of different empirical tests, the dominance of CFA as an empirical test, the use of only one empirical test, and more. These themes are reported in detail below as well as descriptions of unique empirical tests that were used.

Empirical test frequency

The first notable takeaway was the frequency with which empirical tests of discriminant validity are being conducted. In our sample of five journals across two years, only 129 out of 849 articles (15.2%) conducted an empirical test of the distinctiveness of their constructs. In total, only 246 discriminant validity inferences were tested.

Overreliance on CFA

The current empirical practices for justifying discriminant validity of constructs were heavily reliant on CFA. The majority of studies (91.9%) coded used a CFA to test the empirical distinctiveness of their constructs.

We computed descriptive statistics of the main features and fit indices of these CFA models reported in the literature (Table 1). Although not represented in Table 1, item parceling was an unexpected, but frequent, practice in the literature. Thirty-eight studies reported using item parceling to fit their CFA models. An investigation of the effects of item parceling on conclusions of the presence of unitary constructs was outside the bounds of our study, but should be considered by researchers as a variable that could influence the results of their global fit indices and, consequently, their conclusions about discriminant validity inferences.

Notably, Table 1 is an incomplete snapshot of the CFA models being run, because not all researchers chose to report global fit indices or complete descriptions of the alternative CFA models that they ran and compared to their best fitting model. In fact, 15.9% of the articles that reported a CFA reported details for only a subset of CFA models that they claimed to have tested. For example, one article in our sample discussed the best fitting models from their two studies and simply stated, “model comparison tests

revealed that these models fit better than several alternative models in which one or more of the factor correlations were constrained to one." Some studies like this one even failed to report the exact number of models that they ran and chose to only report the number of factors and global fit indices for the chosen model that supported their hypothesis on the discriminant validity of the study constructs.

Single empirical tests

CFA was the dominant empirical test of discriminant validity, and it was often used in isolation. 78% of studies only used information from CFA model comparisons to justify their conclusions about the distinctiveness of their constructs. The use of a single piece of evidence occurred regardless of the empirical test used as evidence. In fact, most studies relied on one empirical test to make a conclusion. Only 14.6% of studies coded in our sample reported more than one empirical test to justify the distinctiveness of their constructs.

Additional empirical tests of discriminant validity

Aside from CFAs, observed correlations, latent variable correlations, exploratory factor analyses (EFA), AVE, and regression-based methods were leveraged to justify the distinctiveness of constructs. Disattenuated observed correlations were never reported, only latent variable correlations between constructs, and latent variable correlations were not reported frequently. Only ten studies out of 246 reported the latent variable correlation.

A common complement to the latent variable correlation, average variance extracted (AVE), was reported in only twelve studies. Studies reporting both the latent variable correlation and AVE values cited Fornell and Larcker's (1981) recommendation

to compare the AVE values for each scale to the squared correlation between latent construct pairs. However, only two studies reported both AVE values and latent variable correlations using Fornell and Larcker's (1981) rule. The remaining ten studies reported AVE values and simply claimed that they were greater than the squared correlation of the latent factors without actually reporting the magnitude of the correlation and/or squared correlation.

The remaining eight studies that reported the latent variable correlation interpreted the magnitude of the correlation to justify the distinctiveness of their constructs. For certain construct pairs with low magnitudes (i.e., < 0.20), this was a straightforward process. For construct pairs with larger latent variable magnitudes, authors often reported additional empirical tests such as a CFA. As an example, one study contained a construct pair that had a latent variable correlation of 0.75. At that magnitude, some researchers would argue against the distinctiveness of the constructs. To build their argument in support of discriminant validity, the authors chose to compute a 95% confidence interval around the latent correlation. They argued that since the confidence interval did not contain 1.0, their discriminant validity inference was supported.

Other studies with large latent variable correlations took a different approach. Six studies recognized that latent correlation magnitudes above 0.70 were questionable in terms of empirical distinctiveness, but argued for the theoretical distinctiveness of their constructs. Therefore, they fit a higher-order one-factor model and treated the individual constructs as subscales. These were some of the few instances where theory was

explicitly highlighted in conjunction with empirical evidence to justify author decisions related to discriminant validity.

The remaining empirical tests used in the literature were observed correlations between measures and regression-based methods. Although we coded the observed correlations between each measure that was tested for distinctiveness (as reported in correlation matrices), only a few articles explicitly mentioned the magnitude of the observed correlation in justifying their decision to treat two constructs as distinct.

Regression-based methods were the least frequent empirical tests conducted and reported in the literature. Only four studies used a regression-based method to justify their decision. Regression based-methods involve any comparison of linear regression models such as discriminant predictive validity or incremental validity.

Overall, results from Study 1 showed that very few researchers are empirically testing the discriminant validity of their constructs. Furthermore, when empirical tests of discriminant validity are conducted, they almost always rely on CFA.

CHAPTER 3: STUDY 2 - MONTE CARLO SIMULATION

Given the aforementioned constraints of CFA and its systematic use in discriminant validity testing, Study 2 explored the consistency with which a CFA supports a 2-factor model when the population data actually reflects a 1-factor model. The study sought to answer the research question: how often do our empirical rules suggest that there are multiple latent constructs when the data was simulated using a congeneric model and how do other features of the data and measures impact that conclusion?

Method

To explore the given research question, a Monte Carlo simulation was programmed using the lavaan package (Rosseel, 2012) in R (R Core Team, 2019) to evaluate the different procedures for assessing discriminant validity. The simulation tested the effectiveness of one of the more commonly used empirical tests of discriminant validity, CFA. First, the simulation generated baseline data from perfect 1-factor models. In organizational research, such a model is implausible. Therefore, we allowed for the latent variable correlation between hypothetical factors to deviate from 1. When the latent variable correlations were less than 1, we were no longer generating a true single factor model. Nevertheless, it was instructive to learn how impactful small deviations in the latent variable correlation could be.

In addition to manipulating the population latent variable correlation, we also manipulated various features of a factor analysis (Table 2), including sample size, magnitude of the factor loadings, and the total number of items within the two scales (e.g., a level of 12 indicates two scales with six items each). We established the levels of

the features to mimic the conditions being reported in the literature that we observed in Study 1. With an average sample size of 284 in the studies coded in the systematic literature review, we simulated results for sample sizes of 200, 350, and 500. The number of scale items condition had three levels (8, 12, 16). These levels were chosen based on the average number of scale items from data collected in Study 1 ($N = 6$). We ran 2,500 iterations for each unique combination of features under each systematic error condition for a total of 325,000 iterations.

After we ran the simulation, we calculated descriptive statistics and ran two logistic regression models to test the influence of four features (sample size, loadings, number of total items, and population latent variable correlation) on the likelihood of concluding that there are two latent factors using the ΔCFI and $\Delta\chi^2$ rules. Relative weights analysis (RWA) was then applied to the two logistic regression models to evaluate which feature(s) were driving the likelihood of concluding that there were two latent factors. Based on these results, we evaluated the accuracy of current methodologies in supporting a one-factor model over a two-factor model.

Results

The results of the Monte Carlo simulation defined the frequency with which various rules support a model with an additional latent factor when compared to a more parsimonious single factor model (Tables 3, 4 and 5). The reliability of the different scales (recorded in the scale 1 raw alpha and scale 2 raw alpha rows) were all acceptable based on commonly accepted cutoffs with values around or above 0.7. The likelihood of rejecting the single factor model is indicted by the performance of the $\Delta\chi^2$ and ΔCFI rules. Their performance was recorded in Tables 3, 4, and 5 as critical scores representing

the proportion of times ΔCFI was greater than 0.01 and the proportion of times that the $\Delta\chi^2$ was greater than four. Looking at the top entry in each of these tables, these results show that when the data is simulated as a single latent factor model with only random measurement error, the $\Delta\chi^2$ and the ΔCFI rules appear to perform very well, even conservatively, by accurately recognizing that these two scales should be considered a single latent factor.

However, the manipulation of the population latent variable correlation to values less than 1.0 elicit a different set of results. These results can be found in the bottom rows of Table 3 for 8-item scales and in the bottom rows of Tables 4 and 5 for 12-item and 16-item scales. Even when the latent variable correlation was as high as 0.95 for 8-item scales, the $\Delta\chi^2$ rule concluded that there are two latent factors 39-99% of the time (Table 3). When the data were generated with a population latent variable of 0.90 for 8-item scales, the $\Delta\chi^2$ rule concluded that there were two latent factors 84-100% of the time (Table 3). Similar results were found for 12-item and 16-item scales (Tables 3-5).

The pattern of results was identical for the ΔCFI rule, although the ΔCFI rule was more conservative than the $\Delta\chi^2$ rule and was not as sensitive to changes in the sample size. When the latent variable correlation was 0.95 for 8-item scales, the ΔCFI rule concluded that there were two latent factors 8-40% of the time (Table 3). When the population latent variable was reduced to 0.90 for 8-item scales, the ΔCFI rule concluded that there were two latent factors 54-99% of the time (Table 3). Similar results were found for 12-item and 16-item scales (Tables 3-5).

These results indicated a clear influence of the latent variable correlation on discriminant validity conclusions, but we built logistic regression models and conducted

relative weights analyses to test our interpretations of the simulation results. Summaries of the logistic regression models can be found in Table 6. The number of items ($b = 0.00$, $p < 0.001$), loadings ($b = 0.78$, $p < 0.001$), sample size ($b = 0.00$, $p < 0.001$), and population latent variable correlation ($b = -2.28$, $p < 0.001$), were all statistically significant, likely due to the sample size of the Monte Carlo simulation. Again, similar results were found in the model with identical features predicting the likelihood of concluding that there are two latent factors using the $\Delta\chi^2$ rule.

Since all of our simulation features were significantly different from zero in our models, we used RWA to identify which features were most important. According to our two logistic regression RWAs (Table 7), the simulation feature with the largest contribution in explaining variance in the likelihood of concluding that there are two latent factors using the ΔCFI rule and $\Delta\chi^2$ rule was the population latent correlation. Respectively, loadings, the total number of scale items, and sample size followed in terms of variable importance, but all had minimal influences on the conclusions drawn from the CFA.

In summary, the population latent variable correlation was a significant predictor in concluding that there was one latent factor and was also shown to be a driver of the variation in the outcome based on the results from the relative weights analysis. This was unsurprising given the pattern of results displayed in Tables 3, 4, and 5. Given these results and the results from Study 1, we consulted experts in Study 3 to generate more defined recommendations for interpreting CFAs, latent variable correlations, and other evidence of discriminant validity.

CHAPTER 4: STUDY 3 - *ORGANIZATIONAL RESEARCH METHODS* (*ORM*) EDITORIAL BOARD SURVEY

The results of Study 1 and Study 2 demonstrate a problem and lack of direction in the reporting and interpretation of CFAs for discriminant validity support. By today's standards, latent variable correlations go largely unreported, but play a large role in discriminant validity testing. Contrarily, observed correlations between two constructs are more frequently reported by organizational researchers. Although we believe in more consistent reporting of the latent variable correlation based on the results of Study 2, no guidelines exist for interpreting these statistics in the context of discriminant validity testing. To address this concern, we strive to provide clarity and specific recommendations on best practices in running and reporting CFAs to avoid construct proliferation. We conducted a survey to collect expert opinions on these recommendations.

Method

Participants

We surveyed the editorial board members for the *Organizational Research Methods* (*ORM*) journal to assess the commonly defined standards for supporting discriminant validity with latent variable correlation values or observed correlation values. This board was selected because of the depth of their expertise in organizational research methods and their familiarity with the organizational science literature. A total of 102 individuals were contacted directly to request their participation in the survey.

Survey procedure

A survey was sent through email to the *ORM* editorial board and we received 30 responses (response rate of 29%). We asked the board six open-ended questions (Appendix B). After collecting responses to the survey, we ran descriptive statistics to determine the expert standards of empirically supporting discriminant validity inferences and the benchmark by which organizational researchers should question the discriminant validity of two proposed constructs.

Results

When asked what empirical evidence or theoretical justification they expect researchers to report in making claims about discriminant validity, there was considerable variability in the specific pieces of theoretical and empirical evidence that were deemed necessary (Table 7). However, there was agreement that reporting the global fit indices of a CFA alone is insufficient. Each respondent listed at least two or more pieces of empirical or theoretical evidence that should be reported in discussions of discriminant validity. As shown in Table 7, latent variable correlations, differential observed correlations with other constructs, and incremental validity were frequently mentioned as empirical evidence that could be reported in addition to CFA. Many board members also argued for the use of theory to interpret or contextualize quantitative results from empirical evidence such as the magnitude of observed and latent correlations.

On average, respondents to our survey indicated that an observed correlation magnitude of 0.58 and a latent variable correlation magnitude of 0.70 would lead them to question inferences of discriminant validity. Nevertheless, they further highlighted various sources of information that could alter their interpretation of these correlational magnitudes. Approximately half of the responses referenced the importance of context,

both theoretical and methodological, in these decisions. Specifically, eleven respondents stated that they would only feel comfortable making conclusions about discriminant validity using the magnitude of an observed correlation if they had an estimate of reliability and/or evidence of method effects that could inflate the observed correlation between measures.

Given the absence of concerns like measurement error at the construct level, there were fewer conditional responses in regards to a latent variable correlation magnitude at which discriminant validity claims should be reevaluated. Still, respondents noted that the magnitude of a latent variable correlation alone would be insufficient in drawing conclusions about the distinctiveness of constructs. Once again, theoretical, methodological, and information from additional empirical tests were said to be critical in interpreting the magnitude of latent variable correlations in decisions about discriminant validity.

The final takeaway from the survey was the warning against empirical cutoffs. Given the many contextual factors described above, members of the editorial board suggested that cutoffs be avoided in making a final conclusion. Nevertheless, certain latent variable correlation magnitudes are more concerning than others, regardless of context, when considering the results found in Study 2.

CHAPTER 5: DISCUSSION

This study sought out to combat construct proliferation through an evaluation of empirical practices for discriminant validity inferences. Construct proliferation is a present threat to the validity of our science; the number of variables in our models is growing (Saylors & Trafimow, 2021). Good measurement practices can guard against the treatment of a single construct as two or more constructs, but these practices are not currently being performed. Study 1, a systematic literature review, recorded the current practices for reporting empirical evidence to justify discriminant validity inferences. One of the main actions taken by researchers in testing a discriminant validity inference is actually inaction. Out of the 849 articles sampled from five top journals across two years, only 129 articles contained empirical tests of discriminant validity inferences. On the rare occasion that empirical tests were conducted, factor analysis, specifically CFA, was the dominant test performed and reported (91.9% of our total sample of 246 studies). Even still, it was often reported incompletely (e.g., only reporting fit statistics for the best fitting model) or done incorrectly in the context of discriminant validity testing (e.g., fitting only one measurement model).

Given the frequency with which researchers are relying on CFA to test the distinctiveness of their constructs, Study 2 focused on the efficacy of this test. A CFA test involves the comparison of global fit indices (e.g., CFI, RMSEA, SRMR, TLI) from nested models. Common cutoffs like 0.01 for ΔCFI and 4 for $\Delta\chi^2$ are frequently cited as support for discriminant validity inferences. Shaffer and colleagues (2016) presented the method of comparing nested CFA models as well as other methods in their review of empirical tests of discriminant validity; however, they also mentioned the considerable

drawbacks and assumptions that are present in such a test. We heeded the warning of Shaffer et al. (2016), tested the ΔCFI and $\Delta\chi^2$ rules, and evaluated their efficacy in making conclusions about the distinctiveness of two constructs. Results showed that CFA tests liberally support construct distinctiveness, even when other sources of empirical evidence would indicate otherwise. When data were simulated from a true one-factor model, the ΔCFI and $\Delta\chi^2$ rules accurately concluded the existence of a single latent factor. But when the latent variable correlation magnitude was reduced slightly to just 0.95 the $\Delta\chi^2$ rule concluded that there are two latent factors 39-99% of the time and the ΔCFI rule concluded that there were two latent factors 8-40% of the time (Table 3). These results indicate that the overwhelmingly popular CFA test is liberally concluding that there are two latent constructs and potentially fueling the issue of construct proliferation in organizational science.

After recognizing shortcomings in the current practices and the evidence against the singular use of CFA to support discriminant validity inferences, Study 3 asked experts, the ORM editorial board members, what researchers should report to justify the distinctiveness of their constructs. We found that current practices for conducting empirical tests in the literature as defined by Study 1 are misaligned with experts' opinions about how and when measures should be considered empirically distinct. Most notably, every respondent advocated for the reporting of two or more pieces of evidence to justify discriminant validity conclusions. Clearly, this is not being demonstrated in the literature since 85.4% of our sample only conducted a single empirical test and theoretical evidence was rarely mentioned. Additionally, some of the most commonly mentioned pieces of empirical evidence that *ORM* editorial board members expect to see

in justifying discriminant validity inferences were absent in the literature. For example, researchers are not exploring or reporting differing patterns of correlations with constructs. Researchers are reporting observed correlations, but only 8% of studies interpreted observed correlation magnitudes in the context of discriminant validity. Moreover, this subset of studies interpreting observed correlation magnitudes are often interpreting the correlations between study variables, not correlations with other measures in the nomological network.

Our survey in Study 3 also asked experts to give correlational magnitudes at which discriminant validity is questionable. On average, experts suggested that if they were presented with a latent variable correlation at or below 0.70, it would make them question the distinctiveness of the two constructs. Their opinions are supported by the results of Study 2. At a population latent variable correlation of 0.70, 100% of the CFAs from our simulation (at varying levels of sample sizes, factor loadings, and total number of items) concluded that there were two latent constructs using the $\Delta\chi^2$ rule. 100% of the CFAs made the same conclusion using the ΔCFI rule. Even Fornell and Larcker's (1981) rule (a more conservation empirical test than the CFA rules) concluded that there were two latent constructs 41 to 100% of the time at a population latent variable correlation of 0.70.

We know that these are three well-known and well-used empirical cutoffs for discriminant validity in the literature, the $\Delta\chi^2$ rule and ΔCFI rule especially, given the results of Study 1. And yet, this pattern of results also shows that it is possible for these empirical rules to support different conclusions in a single scenario. The variability in conclusions from three popular tests in simulation results from Study 2 highlights a

previously unrecognized concern in empirical tests of discriminant validity: nuance. Discriminant validity inferences are not universally supported by a single empirical test at a specified cutoff and it is not guaranteed that all empirical tests will support the same conclusion. Unfortunately, our current guidelines for empirical tests of discriminant validity fail to recognize this reality.

Recommendations

In response to the nuance demonstrated by conversations of discriminant validity inferences and the collective insights from our three studies, we generated specific guidelines to researchers in organizational science when evaluating the distinctiveness of their measures. These recommendations were identified using the aforementioned findings from Study 1 and Study 2 as well as expert opinions from the *ORM* editorial board in Study 3.

Empirically test discriminant validity inferences

First, researchers should increase the practice of empirically testing the discriminant validity of their constructs. Less than half of the articles from the systematic literature review of five *top tier* journals conducted empirical tests of discriminant validity. Without empirical tests of discriminant validity, we can expect the issue of construct proliferation to increase and the accuracy of our models and theories to decrease. Contrarily, rigorous empirical testing of discriminant validity can improve the measurement of organizational science constructs, which is the foundation of our science.

Diversify empirical tests of discriminant validity inferences

We define rigorous empirical testing of discriminant validity as conducting and reporting *multiple* empirical tests that are synthesized and interpreted by the researcher.

Researchers should not rely on a singular empirical test and should diversify the types of empirical tests of discriminant validity. As mentioned previously, only 14.6% of investigations of discriminant validity inferences used more than one empirical test to justify their decision.

Traditional methods like CFA should continue to be used only to provide information regarding the item loadings and overall construct validity of the measures. We recommend that researchers in organizational science avoid the comparison of global fit indices in nested CFA models. As demonstrated in our studies, traditional cutoffs and guidelines for interpreting these model comparisons are frequently leading researchers to conclude that there are two latent constructs when most would argue otherwise. Instead, researchers should compute more insightful statistics like the estimated latent variable correlation and average variance extracted from these CFA models. There are additional empirical tests mentioned by *ORM* editorial board members that can be found in Table 7, but we will focus our conversation on latent variable correlations and AVE since they were evaluated in our Monte Carlo simulation and logistic regression models.

Latent variable correlation. Our Monte Carlo simulation suggested that the latent variable correlation between constructs is one of the main drivers of conclusions of discriminant validity. Results from relative weights analyses of two logistic regression models predicting the likelihood of concluding that there are two latent factors showed that 79% of the variance in ΔCFI rule conclusions and 78% of the variance in $\Delta\chi^2$ rule conclusions are driven by the magnitude of the population latent variable correlation. Moreover, *ORM* editorial board members expressed the importance of the latent variable correlation in drawing conclusions about discriminant validity inferences. In spite of the

importance of the latent variable correlation in discriminant validity tests, as acknowledged by board members at *ORM* and the simulation results, results from Study 1 showed that very few studies are reporting latent variable correlations ($N = 10$).

Instead of interpreting global fit indices, researchers should report and interpret the magnitude of the latent variable correlations amongst constructs of interest. The magnitude of 0.70 suggested by *ORM* board members should guide a researcher in their conclusion about the distinctiveness of the constructs, but not be used as an explicit cutoff. Theory and additional tests should always be reported in addition to latent variable correlations. As the latent variable correlation magnitude exceeds 0.70, the burden of proof necessary to conclude that there are two latent constructs should increase accordingly. The integration of theoretical and empirical evidence is discussed in greater detail below.

Average variance extracted. One test we recommend reporting in addition to the latent variable correlation is the AVE for each construct. Results from our studies further supported the use of Fornell and Larcker's (1981) rule of thumb that compares the AVE to the squared latent variable correlation. We evaluated the performance of this rule in our Monte Carlo simulation as well. When the population latent variable correlation was 0.8, the AVE values for each individual scale were greater than the squared latent correlation 1-40% of the time. When the population latent variable correlation was reduced to 0.7, Fornell and Larcker's (1981) rule concluded that there were two distinct constructs 42-100% of the time.

These results indicate that the comparison of AVE values to the shared variance between constructs is less likely to conclude that there are two distinct constructs than the

frequently used CFA rules (i.e., the $\Delta\chi^2$ and ΔCFI). We are not suggesting that the comparison of AVE values and shared construct variance should replace other existing tests, rather that this empirical test can be a complement to the evidence presented by latent variable correlations and other empirical tests of discriminant validity.

Leverage both empirical tests and theoretical justification of discriminant validity

Finally, when researchers conduct empirical tests of discriminant validity, they should report both empirical and theoretical evidence. Although the focus of this study was on empirical tests of discriminant validity, theoretical justification is another necessary and key element of discriminant validity inference testing. Although we did not explicitly code for theoretical justification in our systematic literature, it was noted by our coding team as being consistently absent in discussions about discriminant validity inferences. Instead, researchers are interpreting their empirical findings in accordance with traditional cutoffs that identify a best-fitting CFA model. But theoretical justification enables researchers to explore to nuances in discriminant validity inferences that cutoff values would not allow. As mentioned previously, empirical tests do not always elicit complementary findings.

Theoretical justification is necessary at multiple stages. First, researchers should follow Podsackoff and colleagues' (2016) guidance on creating conceptually distinct construct definitions. Researchers should consider these construct definitions in identifying constructs that should be empirically tested for discriminant validity. In Study 1 results, many studies chose to test their full measurement model rather than run empirical tests on specific construct pairs that were theoretically similar. This led to the comparison of large models with up to 19 factors. Although the number of factors to

include in factor analyses is highly dependent on the study, researchers should only empirically test constructs that have questionable distinctiveness or are in the same nomological network rather than the full measurement model.

The second opportunity to integrate theory into the empirical testing of discriminant validity hypotheses is during the interpretation of empirical evidence. Consider again the examples from Study 1 where researchers compared three-factor and one-factor models. The one-factor model demonstrated better fit, but the researchers argued that their scales were still conceptually distinct. Therefore, they chose to move forward with a one-factor higher order model in order to recognize the empirical evidence of the presence of a single latent construct, but also retain the theoretical distinction of their scales.

Theory can also be used to interpret the magnitude of both observed and latent variable correlations. A true one-factor CFA model has a latent factor correlation of 1.0. Any magnitude below that value is left up to interpretation in our social science. The interpretation of correlational magnitudes (both observed and latent) is reliant on our construct definitions and other theoretical pieces of evidence. As mentioned by survey respondents in Study 3, there are constructs in our literature that share a lot of conceptual similarity, but we treat them as theoretically distinct constructs. In examples like these, we would expect to see latent variable correlation magnitudes at or even above 0.70. This is less concerning than scenarios with construct pairs that should not share high correlational magnitudes or have less conceptual overlap; however, researchers must show that the latent variable correlations and observed correlations reflect theoretical and conceptual evidence, regardless of their magnitudes.

In summary, theory guides the creation of conceptual distinctiveness and the interpretation of empirical evidence. Again, this interpretation is especially important when a researcher appropriately conducts multiple empirical tests of discriminant validity that give contradictory results. This paper was focused on empirical tests of validity, but researchers ultimately cannot conduct rigorous empirical tests of discriminant validity without the application of theory.

Limitations and Future Research

This paper focused on discriminant validity tests at a single level of analysis. Future research should investigate any key differences in testing and justifying discriminant validity inferences in multilevel models.

CHAPTER 6: CONCLUSION

Extant literature offers guidance on numerous empirical tests of discriminant validity; nevertheless, considerable assumptions are made regarding the ability of these tests to make accurate inferences and there are very few explicit guidelines on how to best conduct empirical tests of discriminant validity. This paper expanded past work on empirical tests of discriminant validity by defining current practices for testing the distinctiveness of constructs, evaluating the efficacy of these practices, and surveying experts to define new best practice guidelines for conducting empirical tests of discriminant validity inferences. The vast majority of researchers currently compare nested CFA models to support the distinctiveness of their constructs, but we found that these tests fail to accurately identify the presence of one or two latent constructs. Instead of conducting CFA, we suggest that organizational researchers follow the advice of *ORM* board members to use multiple pieces of empirical evidence to justify their discriminant validity inferences. Specifically, we identified latent variable correlations and Fornell and Larcker's (1981) AVE rule as strong empirical tests. Finally, theoretical justification is a necessary component of discriminant validity tests that should be used in conjunction with empirical evidence to make an ultimate conclusion about the existence of one or more latent constructs.

Table 1

Summary of CFA models reported in JAP, JOM, AMJ, SMJ, and PPsych from 2018-2019

Feature	***M***
Sample size	222
Number of factors	4.4
Number of scale items	6.22
Coefficient alpha	0.87
Number of nested CFA models compared	2.78

Note. $N = 626$ CFA models.

Table 2

Monte Carlo simulation features

Feature	Levels
Sample size	200, 350, 500
Factor loadings	0.7, 0.8
Number of total items	8, 12, 16
Population LVC	1.0, 0.95, 0.90, 0.85, 0.80, 0.75, 0.70

Table 3

Mean results and two factor likelihood for 8-item scale CFAs

		λ = .7			λ = .8		
	Sample size	200	350	500	200	350	500
Population model LVC = 1	Latent variable correlation	0.98	0.98	0.99	0.99	0.99	0.99
	Scale score correlation	0.78	0.78	0.78	0.87	0.87	0.87
	Scale 1 raw alpha	0.80	0.80	0.80	0.88	0.88	0.88
	Scale 2 raw alpha	0.80	0.80	0.80	0.88	0.88	0.88
	Critical score (ΔCFI)	0.01	0.00	0.00	0.00	0.00	0.00
	Critical score ($\Delta\chi^2$)	0.05	0.05	0.04	0.05	0.05	0.04
	Critical score (AVE > ρ^2)	0.00	0.00	0.00	0.00	0.00	0.00
Population model LVC = 0.95	Latent variable correlation	0.95	0.95	0.95	0.95	0.95	0.95
	Scale score correlation	0.75	0.75	0.75	0.83	0.83	0.83
	Scale 1 raw alpha	0.79	0.79	0.79	0.87	0.88	0.88
	Scale 2 raw alpha	0.79	0.79	0.79	0.88	0.88	0.88
	Critical score (ΔCFI)	0.13	0.09	0.07	0.26	0.26	0.24
	Critical score ($\Delta\chi^2$)	0.39	0.59	0.72	0.80	0.96	0.99
	Critical score (AVE > ρ^2)	0.00	0.00	0.00	0.00	0.00	0.00
Population model LVC = 0.90	Latent variable correlation	0.90	0.90	0.90	0.90	0.90	0.90
	Scale score correlation	0.71	0.71	0.71	0.79	0.79	0.79
	Scale 1 raw alpha	0.79	0.79	0.79	0.88	0.88	0.88
	Scale 2 raw alpha	0.79	0.79	0.79	0.88	0.88	0.88
	Critical score (ΔCFI)	0.54	0.66	0.72	0.91	0.98	0.99
	Critical score ($\Delta\chi^2$)	0.84	0.98	0.99	1.00	1.00	1.00
	Critical score (AVE > ρ^2)	0.00	0.00	0.00	0.00	0.00	0.00
Population model LVC = 0.85	Latent variable correlation	0.85	0.85	0.85	0.85	0.85	0.85
	Scale score correlation	0.67	0.67	0.67	0.74	0.74	0.74
	Scale 1 raw alpha	0.79	0.79	0.79	0.88	0.88	0.88
	Scale 2 raw alpha	0.79	0.79	0.79	0.88	0.88	0.88
	Critical score (ΔCFI)	0.88	0.97	0.97	1.00	1.00	1.00
	Critical score ($\Delta\chi^2$)	0.99	1.00	1.00	1.00	1.00	1.00
	Critical score (AVE > ρ^2)	0.00	0.00	0.00	0.04	0.02	0.00

Population model LVC = 0.80	Latent variable correlation	0.80	0.80	0.80	0.80	0.80	0.80
	Scale score correlation	0.63	0.63	0.63	0.70	0.70	0.70
	Scale 1 raw alpha	0.79	0.79	0.79	0.88	0.88	0.88
	Scale 2 raw alpha	0.79	0.79	0.79	0.88	0.88	0.88
	Critical score (ΔCFI)	0.98	1.00	1.00	1.00	1.00	1.00
	Critical score ($\Delta\chi^2$)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score (AVE > ρ^2)	0.01	0.00	0.00	0.42	0.43	0.40
Population model LVC = 0.75	Latent variable correlation	0.75	0.75	0.75	0.75	0.75	0.75
	Scale score correlation	0.60	0.59	0.59	0.66	0.66	0.66
	Scale 1 raw alpha	0.79	0.79	0.79	0.88	0.88	0.88
	Scale 2 raw alpha	0.79	0.79	0.79	0.88	0.88	0.88
	Critical score (ΔCFI)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score ($\Delta\chi^2$)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score (AVE > ρ^2)	0.13	0.07	0.06	0.85	0.93	0.96
Population model LVC = 0.70	Latent variable correlation	0.70	0.70	0.70	0.70	0.70	0.70
	Scale score correlation	0.56	0.55	0.56	0.61	0.61	0.61
	Scale 1 raw alpha	0.79	0.79	0.79	0.88	0.88	0.88
	Scale 2 raw alpha	0.79	0.79	0.79	0.88	0.88	0.88
	Critical score (ΔCFI)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score ($\Delta\chi^2$)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score (AVE > ρ^2)	0.42	0.42	0.41	0.98	1.00	1.00

Table 4

Mean results and two factor likelihood for 12-item scale CFAs

		$\lambda = .7$			$\lambda = .8$		
	Sample size	200	350	500	200	350	500
Population model LVC = 1	Latent variable correlation	0.99	0.99	0.99	0.99	0.99	0.99
	Scale score correlation	0.84	0.84	0.84	0.91	0.91	0.91
	Scale 1 raw alpha	0.85	0.85	0.85	0.91	0.91	0.91
	Scale 2 raw alpha	0.85	0.85	0.85	0.91	0.91	0.91
	Critical score (ΔCFI)	0.00	0.00	0.00	0.00	0.00	0.00
	Critical score ($\Delta\chi^2$)	0.03	0.03	0.05	0.05	0.05	0.05
	Critical score (AVE > ρ^2)	0.00	0.00	0.00	0.00	0.00	0.00
Population model LVC = 0.95	Latent variable correlation	0.95	0.95	0.95	0.95	0.95	0.95
	Scale score correlation	0.81	0.81	0.81	0.87	0.87	0.87
	Scale 1 raw alpha	0.85	0.85	0.85	0.91	0.91	0.91
	Scale 2 raw alpha	0.85	0.85	0.85	0.91	0.91	0.91
	Critical score (ΔCFI)	0.13	0.11	0.08	0.39	0.43	0.45
	Critical score ($\Delta\chi^2$)	0.67	0.88	0.96	0.98	1.00	1.00
	Critical score (AVE > ρ^2)	0.00	0.00	0.00	0.00	0.00	0.00
Population model LVC = 0.90	Latent variable correlation	0.90	0.90	0.90	0.90	0.90	0.90
	Scale score correlation	0.77	0.77	0.77	0.82	0.82	0.82
	Scale 1 raw alpha	0.85	0.85	0.85	0.91	0.91	0.91
	Scale 2 raw alpha	0.85	0.85	0.85	0.91	0.91	0.91
	Critical score (ΔCFI)	0.73	0.86	0.92	0.99	1.00	1.00
	Critical score ($\Delta\chi^2$)	0.99	1.00	1.00	1.00	1.00	1.00
	Critical score (AVE > ρ^2)	0.00	0.00	0.00	0.00	0.00	0.00
Population model LVC = 0.85	Latent variable correlation	0.85	0.85	0.85	0.85	0.85	0.85
	Scale score correlation	0.72	0.72	0.72	0.78	0.78	0.78
	Scale 1 raw alpha	0.85	0.85	0.85	0.91	0.91	0.91
	Scale 2 raw alpha	0.85	0.85	0.85	0.91	0.91	0.91
	Critical score (ΔCFI)	0.98	1.00	1.00	1.00	1.00	1.00
	Critical score ($\Delta\chi^2$)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score (AVE > ρ^2)	0.00	0.00	0.00	0.00	0.00	0.00

Population model LVC = 0.80	Latent variable correlation	0.80	0.80	0.80	0.80	0.80	0.80
	Scale score correlation	0.68	0.68	0.68	0.73	0.73	0.73
	Scale 1 raw alpha	0.85	0.85	0.85	0.91	0.91	0.91
	Scale 2 raw alpha	0.85	0.85	0.85	0.91	0.91	0.91
	Critical score (ΔCFI)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score ($\Delta\chi^2$)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score (AVE > ρ^2)	0.00	0.00	0.00	0.40	0.41	0.40
Population model LVC = 0.75	Latent variable correlation	0.75	0.75	0.75	0.75	0.75	0.75
	Scale score correlation	0.64	0.64	0.64	0.69	0.69	0.69
	Scale 1 raw alpha	0.85	0.85	0.85	0.91	0.91	0.91
	Scale 2 raw alpha	0.85	0.85	0.85	0.91	0.91	0.91
	Critical score (ΔCFI)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score ($\Delta\chi^2$)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score (AVE > ρ^2)	0.09	0.06	0.03	0.89	0.96	0.98
Population model LVC = 0.70	Latent variable correlation	0.70	0.70	0.70	0.70	0.70	0.70
	Scale score correlation	0.59	0.60	0.60	0.64	0.64	0.64
	Scale 1 raw alpha	0.85	0.85	0.85	0.91	0.91	0.91
	Scale 2 raw alpha	0.85	0.85	0.85	0.91	0.91	0.91
	Critical score (ΔCFI)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score ($\Delta\chi^2$)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score (AVE > ρ^2)	0.44	0.43	0.41	1.00	1.00	1.00

Table 5

Mean results and two factor likelihood for 16-item scale CFAs

		λ = .7			λ = .8		
	Sample size	200	350	500	200	350	500
Population model LVC = 1	Latent variable correlation	0.99	0.99	0.99	0.99	1.00	1.00
	Scale score correlation	0.88	0.88	0.88	0.93	0.93	0.93
	Scale 1 raw alpha	0.88	0.89	0.88	0.93	0.93	0.93
	Scale 2 raw alpha	0.88	0.89	0.88	0.93	0.93	0.93
	Critical score (ΔCFI)	0.00	0.00	0.00	0.00	0.00	0.00
	Critical score ($\Delta\chi^2$)	0.05	0.04	0.04	0.05	0.05	0.04
	Critical score (AVE > ρ^2)	0.00	0.00	0.00	0.00	0.00	0.00
Population model LVC = 0.95	Latent variable correlation	0.95	0.95	0.95	0.95	0.95	0.95
	Scale score correlation	0.84	0.84	0.84	0.89	0.89	0.89
	Scale 1 raw alpha	0.88	0.88	0.88	0.93	0.93	0.93
	Scale 2 raw alpha	0.88	0.88	0.88	0.93	0.93	0.93
	Critical score (ΔCFI)	0.18	0.15	0.11	0.54	0.63	0.68
	Critical score ($\Delta\chi^2$)	0.87	0.98	1.00	1.00	1.00	1.00
	Critical score (AVE > ρ^2)	0.00	0.00	0.00	0.00	0.00	0.00
Population model LVC = 0.90	Latent variable correlation	0.90	0.90	0.90	0.90	0.90	0.90
	Scale score correlation	0.80	0.80	0.80	0.84	0.84	0.84
	Scale 1 raw alpha	0.88	0.88	0.88	0.93	0.93	0.93
	Scale 2 raw alpha	0.88	0.88	0.88	0.93	0.93	0.93
	Critical score (ΔCFI)	0.87	0.95	0.98	1.00	1.00	1.00
	Critical score ($\Delta\chi^2$)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score (AVE > ρ^2)	0.00	0.00	0.00	0.00	0.00	0.00
Population model LVC = 0.85	Latent variable correlation	0.85	0.85	0.85	0.85	0.85	0.85
	Scale score correlation	0.75	0.75	0.75	0.79	0.79	0.79
	Scale 1 raw alpha	0.88	0.88	0.88	0.93	0.93	0.93
	Scale 2 raw alpha	0.88	0.88	0.88	0.93	0.93	0.93
	Critical score (ΔCFI)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score ($\Delta\chi^2$)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score (AVE > ρ^2)	0.00	0.00	0.00	0.01	0.00	0.00

Population model LVC = 0.80	Latent variable correlation	0.80	0.80	0.80	0.80	0.80	0.80
	Scale score correlation	0.71	0.71	0.71	0.75	0.75	0.75
	Scale 1 raw alpha	0.88	0.88	0.88	0.93	0.93	0.93
	Scale 2 raw alpha	0.88	0.88	0.88	0.93	0.93	0.93
	Critical score (ΔCFI)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score ($\Delta\chi^2$)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score (AVE > ρ^2)	0.00	0.00	0.00	0.39	0.41	0.41
Population model LVC = 0.75	Latent variable correlation	0.75	0.75	0.75	0.75	0.75	0.75
	Scale score correlation	0.66	0.66	0.66	0.70	0.70	0.70
	Scale 1 raw alpha	0.88	0.88	0.88	0.93	0.93	0.93
	Scale 2 raw alpha	0.88	0.88	0.88	0.93	0.93	0.93
	Critical score (ΔCFI)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score ($\Delta\chi^2$)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score (AVE > ρ^2)	0.07	0.03	0.01	0.90	0.97	0.99
Population model LVC = 0.70	Latent variable correlation	0.70	0.70	0.70	0.69	0.70	0.70
	Scale score correlation	0.62	0.62	0.62	0.64	0.65	0.65
	Scale 1 raw alpha	0.88	0.88	0.88	0.93	0.93	0.93
	Scale 2 raw alpha	0.88	0.88	0.88	0.93	0.93	0.93
	Critical score (ΔCFI)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score ($\Delta\chi^2$)	1.00	1.00	1.00	1.00	1.00	1.00
	Critical score (AVE > ρ^2)	0.42	0.41	0.40	1.00	1.00	1.00

Table 6

Logistic regression model summaries

$\hat{y}$	Intercept	Population latent variable correlation	Loadings	Sample size	Total number of scale items
1. ΔCFI > 0.01	2.85***	-3.29***	0.78***	0.00***	0.00***
2. $\Delta\chi^2 > 4$	2.43***	-2.25***	0.34***	0.00***	0.00***
3. AVE > ρ^2	0.22***	-2.34***	2.62***	0.00***	0.00**

Note. *** $p < 0.001$. ** $p < 0.01$.

Table 7

Rescaled relative weights

$\hat{y}$	Population latent variable correlation	Loadings	Sample Size	Total Number of Scale Items
1. ΔCFI > 0.01	0.98	0.02	0.00	0.01
2. $\Delta\chi^2 > 4$	0.99	0.00	0.00	0.00
3. AVE > ρ^2	0.81	0.19	0.00	0.00

Note. The weights do not sum to 1 due to rounding.

Table 8

ORM editorial board survey: Expected discriminant validity justification

Evidence	***N***
Clear construct definitions	5
Theoretical explanation of distinctiveness	13
Observed correlation between constructs	9
Differential observed relationships	10
Latent variable correlation	11
EFA	6
CFA	20
AVE	1
HTMT	1
Examination of common method bias	1
Evidence of a distinct set of antecedents or differential relationships with antecedents	2
Increment validity	4
Discriminant predictive validity	1

Note. $N = 30$ participants.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Psychological Association. (2018). *Principles for the Validation and Use of Personnel Selection Procedures* (Fifth Edition, pp. 1–60).

Antonakis, J. (2004). On why "emotional intelligence" will not predict leadership effectiveness beyond IQ or the "big five": An extension and rejoinder. *Organizational Analysis, 12*, 171-182.

Antonakis, J., & Dietz, J. (2010). Emotional intelligence: On definitions, neuroscience, and marshmallows. *Industrial and Organizational Psychology, 3*. 165-170. http://dx.doi.org/1754-9426/10

Ashkanasy, N.M., & Daus, C.S. (2005). Rumors of the death of emotional intelligence in organizational behavior are vastly exaggerated. *Journal of Organizational Behavior, 26*. 441-452. http://dx.doi.org/10.1002/job.320.

Bagozzi, R. P. (1981). Evaluating structural equation models with unobservable variables and measurement error: A comment. *Journal of Marketing Research, 18*, 375-381.

Binning, J.F., & Barret, G.V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*(3). 478-494.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Davis, G. F. (2015). Editorial essay: What is organizational research for? *Administrative Science Quarterly, 60*(2). 179-188. http://dx.doi.org/10.1177/0001839215585725

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*, 39–50.

Hambrick, D.C. (2007). The field of management's devotion to theory: Too much of a good thing? *The Academy of Management Journal, 50*(6). 1346-1352.

Harter, J. K., & Schmidt, F. L. (2008). Conceptual versus empirical distinctions among constructs: Implications for discriminant validity. *Industrial and Organizational Psychology*, *1*(1), 36–39. https://doi.org/10.1111/j.1754-9434.2007.00004.x

Le, H., Schmidt, F. L., Harter, J. K., & Lauver, K. J. (2010). The problem of empirical redundancy of constructs in organizational research: An empirical investigation. *Organizational Behavior and Human Decision Processes*, *112*(2), 112–125. https://doi.org/10.1016/j.obhdp.2010.02.003

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

Marsh, H. W., Hau, K.T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*(2), 181–220. https://doi.org/10.1207/s15327906mbr3302_1

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, *5*(2), 159–168. https://doi.org/10.1207/s15327574ijt0502_4

Podsakoff, P.M., MacKenzie, S.B., & Podsakoff, N.P. (2016). Recommendations for creating better concept definitions in the organizational, behavioral, and social sciences. *Organizational Research Methods, 19*(2). 159-203. https//doi.org/10.1177/1094428115624965

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1-36. http://www.jstatsoft.org/v48/i02/.

Sackett, P. R., Putka, D. J., & McCloy, R. A. (2012). The concept of validity and the process of validation. In N. Schmitt (Ed.), *Oxford handbook of assessment and selection*. Oxford, UK: Oxford University Press.

Saylors, R., & Trafimow, D. (2021). Why the Increasing Use of Complex Causal Models Is a Problem: On the Danger Sophisticated Theoretical Narratives Pose to Truth. *Organizational Research Methods*, *24*(3), 616–629. https://doi.org/10.1177/1094428119893452

Shaffer, J. A., DeGeest, D., & Li, A. (2016). Tackling the problem of construct proliferation: A guide to assessing the discriminant validity of conceptually related constructs. *Organizational Research Methods*, *19*(1), 80–110.

Appendix A: Literature Coding Questions and Keywords

Coding Questions

- General Article Information
    - Article Title
    - Author Names
    - Journal Name
    - Journal Year
    - Name of the coder
- What models are they testing? (ex: 3 factor to 2 factor; 2 factor to 1 factor)
- How many latent constructs?
- Sample size?
- Did they report each measure's number of items and reliability? What are those values?
- What are the observed correlations between measures?
- Do they report the latent variable correlations? What are those values?
- What fit indices do they report? Insert the names and values of the reported fit indices for the selected model.
- Did they compare models' change in fit indices? Insert the reported fit indices for the comparison of best alternative model and the selected model.
- Did they report any other relevant statistics and what were their values?
- What evidence do they ultimately use to justify their claim?
- What did they ultimately decide based on their reported statistics? Why did they decide this action?

- Is there anything else you want to share or note about this article?

Keywords from sample coding in the Journal of Applied Psychology (2018)

Keywords
Confirmatory factor analysis or CFA
Confirmatory factor analyses
Confirmatory
Chi Square
Comparative fit index or CFI
Tucker-Lewis Index or TLI
Root mean square error or RMSEA
Standardized root mean square residual or SRMR
Discriminant validity

Appendix B: *ORM* Board Survey Questions

1. If you were reviewing a manuscript and the authors were trying to demonstrate that two measures were distinct, what empirical evidence and/or theoretical justification would you want them to provide to support this claim?
2. Assume the authors present the observed correlation between the two measures that they argue are distinct. At what magnitude of the observed correlation would you question the discriminant validity?
    a. Do you have any additional thoughts or comments you want to share regarding the question above?
3. Assume you are presented with the latent correlation between the two measures. At what magnitude of the latent correlation would you question the discriminant validity?
    a. Do you have any additional thoughts or comments you want to share regarding the question above?
4. Do you have any general thoughts about the current practices for empirically justifying discriminant validity inferences that you would like to share with us?